frontiers in Computer Science

ORIGINAL RESEARCH
published: 22 March 2022
doi: 10.3389/fcomp.2022.796866

Outer Product-Based Fusion of Smartwatch Sensor Data for Human Activity Recognition

Adria Mallol-Ragolta[1]*, *Anastasia Semertzidou*[1], *Maria Pateraki*[2,3] and *Björn Schuller*[1,4]

[1] EIHW – Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Augsburg, Germany, [2] Institute of Computer Science, Foundation of Research and Technology – Hellas, Heraklion, Greece, [3] School of Rural, Surveying and Geoinformatics Engineering, National Technical University of Athens, Athens, Greece, [4] GLAM – Group on Language, Audio, & Music, Imperial College London, London, United Kingdom

The advent of IoT devices in combination with Human Activity Recognition (HAR) technologies can contribute to battle with sedentariness by continuously monitoring the users' daily activities. With this information, autonomous systems could detect users' physical weaknesses and plan personalized training routines to improve them. This work investigates the multimodal fusion of smartwatch sensor data for HAR. Specifically, we exploit pedometer, heart rate, and accelerometer information to train unimodal and multimodal models for the task at hand. The models are trained end-to-end, and we compare the performance of dedicated Recurrent Neural Network-based (RNN) and Convolutional Neural Network-based (CNN) architectures to extract deep learnt representations from the input modalities. To fuse the embedded representations when training the multimodal models, we investigate a concatenation-based and an outer product-based approach. This work explores the harAGE dataset, a new dataset for HAR collected using a Garmin Vivoactive 3 device with more than 17 h of data. Our best models obtain an Unweighted Average Recall (UAR) of 95.6, 69.5, and 60.8 % when tackling the task as a 2-class, 7-class, and 10-class classification problem, respectively. These performances are obtained using multimodal models that fuse the embedded representations extracted with dedicated CNN-based architectures from the pedometer, heart rate, and accelerometer modalities. The concatenation-based fusion scores the highest UAR in the 2-class classification problem, while the outer product-based fusion obtains the best performances in the 7-class and the 10-class classification problems.

Keywords: artificial intelligence, human activity recognition, multimodal fusion, ubiquitous computing, smartwatch sensor data

OPEN ACCESS

Edited by:
Pekka Siirtola,
University of Oulu, Finland

Reviewed by:
Saeed Hamood Alsamhi,
Ibb University, Yemen
Martin Gjoreski,
University of Italian Switzerland, Switzerland

***Correspondence:**
Adria Mallol-Ragolta
adria.mallol-ragolta@informatik.uni-augsburg.de

Specialty section:
This article was submitted to Mobile and Ubiquitous Computing, a section of the journal Frontiers in Computer Science

Received: *17 November 2021*
Accepted: *18 February 2022*
Published: *22 March 2022*

Citation:
Mallol-Ragolta A, Semertzidou A, Pateraki M and Schuller B (2022) Outer Product-Based Fusion of Smartwatch Sensor Data for Human Activity Recognition. Front. Comput. Sci. 4:796866. doi: 10.3389/fcomp.2022.796866

1. INTRODUCTION

According to the *World Health Organization* (WHO), physical inactivity is a serious public health concern with serious implications in people's health, as it can be a risk factor for diabetes, depression, high blood pressure, or obesity. Physical activity is beneficial not only for physical health, but also for wellbeing (Fox, 1999; Penedo and Dahn, 2005). Hence, there is a need to develop new, digital, and personalized tools that engage their users to exercise with the goal to have a more active, and healthier life. The research performed on the field of *Human Activity Recognition* (HAR) can contribute to achieve this goal. This field of knowledge aims to develop technologies able to

recognize and, therefore, monitor the activities that users do. The exploitation of this information has a wide range of applications in many different domains, such as healthcare, fitness, athletics, elderly care, security, or entertainment.

Commercial smartphones are equipped with embedded sensors, including accelerometers and gyroscopes, which make them suitable to recognize human activities (Khan et al., 2010; Bayat et al., 2014; Chen and Shen, 2017). Previous works in the literature explored different machine learning techniques, such as hidden Markov models (Ronao and Cho, 2014), unsupervised learning (Kwon et al., 2014), and deep learning (Ronao and Cho, 2016; Hassan et al., 2018) for HAR using smartphone sensor data. Smartwatches are a high-potential device for this task as well (Weiss et al., 2016; Shahmohammadi et al., 2017; Mekruksavanich and Jitpattanakul, 2020) because of their market penetration in society, which is increasing every year, and the embedded sensors they contain, which can be used to retrieve pedometer, photoplethysmographic, and accelerometer measurements (Lara et al., 2012). Furthermore, their location on the users' wrist seems advantageous to capture human activities.

This work investigates the multimodal fusion of pedometer, heart rate, and accelerometer information to train end-to-end models for HAR. One of the goals of this work is to determine which modalities are more suitable to be fused for the task at hand. Based on the tensor fusion layer presented by Zadeh et al. (2017), we propose using an outer product-based approach to fuse the embedded representations of the input modalities. The performance of this approach is compared with a concatenation-based approach, which we use as a baseline. The embedded representations of the input modalities are learnt using dedicated *Recurrent Neural Network*-based (RNN) or *Convolutional Neural Network*-based (CNN) architectures. Our experiments explore the harAGE dataset, a new smartwatch-based HAR dataset collected using a Garmin Vivoactive 3 device. The dataset contains more than 17 h of data from 19 participants while lying, sitting, standing, washing hands, walking, running, climbing stairs, doing strength and flexibility workout activities, and cycling.

The rest of this article is laid out as follows. Section 2 first highlights relevant related works in the literature. Section 3 presents the dataset employed, while Section 4 describes the methodology followed. Section 5 analyzes the results obtained from the experiments performed, and Section 6 concludes this article and suggests some future work directions.

2. RELATED WORK

The problem of HAR is an active topic in the research community. A large body of knowledge has tackled the problem from a computer vision perspective (Khaire et al., 2018; Qi et al., 2018), exploiting color, depth, and even skeletal information. We can consider these as passive approaches, as they require cameras overseeing the scene to perform inferences. On the other side, we can consider as active approaches those that use body-worn sensors for recognizing human activities. In this case, the sensors themselves experience the activities, and, therefore, the sensor measurements can be directly used to infer them. Research on this topic has been conducted using dedicated heart rate sensors (Tapia et al., 2007), inertial/magnetic sensors (Altun and Barshan, 2010), or accelerometer sensors (Lin et al., 2018).

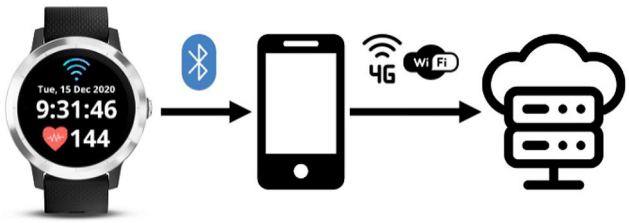

FIGURE 1 | Diagram illustrating the infrastructure implemented to collect the harAGE dataset and the technologies used to transfer the data between the involved nodes. The front-end design of the customized smartwatch app is depicted on the left.

A wide range of sensors are embedded in consumer, smart devices nowadays, including smartphones and smartwatches. Their high penetration in society has motivated the use of data collected with such devices for HAR purposes (Ahmed et al., 2020; Ashry et al., 2020; Mekruksavanich and Jitpattanakul, 2020; Wan et al., 2020). From a user-centered perspective, the field of HAR has traditionally focused on recognizing the activities individuals do. Nevertheless, recent works are considering the problem from a *Multi-user Activity Recognition* (MAR) perspective, which addresses the activities that a group of individuals do to achieve a common goal (Li et al., 2020).

Multimodal approaches have been used in a wide variety of problems and applications to complement and enrich the information embedded in a single modality. Different fusion techniques, from simple to complex, have been explored for this purpose. Examples of simple fusion techniques include the element-wise sum or product of the features extracted from different modalities, or even their simple concatenation. Among the more complex techniques, researchers have investigated circulant fusion (Wu and Han, 2018), gated fusion (Kim et al., 2018), memory (Priyasad et al., 2021), graph neural networks (Holzinger et al., 2021), and even transformers (Prakash et al., 2021).

3. DATASET

This work explores the first version of harAGE: a new smartwatch-based dataset for HAR collected using a customized smartwatch app running on a Garmin Vivoactive 3 device (Mallol-Ragolta et al., 2021). The app reads the accelerometer, the heart rate, and the pedometer information available from the built-in embedded sensors. While the accelerometer information is sensed at 25 Hz, the sampling rate of the heart rate and the steps information is 1 Hz. The back-end of the smartwatch app encapsulates the data into a JSON message, which is sent in close to real-time into a customized, encrypted, and secure server via the Internet using the HTTPS protocol (cf. **Figure 1**).

TABLE 1 | Summary of the activities included in the harAGE dataset, the number of participants collected for each activity, and the amount of data available time-wise.

Activity	Participants	Duration (HH):MM:SS
Resting	19	1:25:24
Lying	19	1:39:01
Sitting	18	1:31:25
Standing	18	1:35:51
Washing hands	18	53:40
Walking	18	2:23:59
Running	16	1:58:28
Stairs climbing	18	2:17:23
Strength Workout	18	53:05
Flexibility Workout	18	56:50
Cycling	13	1:36:40
$\sum$	19	17:11:46

The recruited participants followed a protocol especially designed for the collection of the harAGE dataset. The participants started with a resting phase during 5 min to collect their heart rate at rest, avoiding stressors and external stimuli. This measurement can be used as the baseline heart rate for each individual participant. Then, they performed a sequence of static activities including lying, sitting, and standing. These three activities were performed twice: first without moving, and then allowing reasonable free movements. Each one of these activities was performed during 3 min. Next, we asked participants to simulate washing their hands, without running water, also for 3 min. Although this activity was rarely included in previous HAR datasets found in the literature, the current pandemic context and the favorable placement of the smartwatch in the participants' wrist motivated its inclusion in the data collection protocol.

The following dynamic activities were included next in the protocol: walking, running, climbing stairs (both upstairs and downstairs), and cycling. Furthermore, each one of these activities was performed three times at low, moderate, and high intensities during 3 min each. Intensity levels are subjective, as these depend on several factors, such as the previous physical condition of the participants. Thus, to capture this variability in our dataset, we relied on the participants themselves to set their own thresholds for each intensity level. Before the cycling set of activities, we incorporated a set of workout activities in the protocol. These activities included two sets of strength workout activities (squats and arm raising exercises), and two sets of flexibility workout activities (shoulder roll and wrist stretching exercises). These four activities were performed for 1.5 min each.

To guarantee the safety measures against the COVID-19 pandemic, the dataset was mainly collected outdoors. This scenario posed a challenge, as the data transfer between the smartwatch and the server when the participants were outdoors was performed via the 4G connection of the smartphone with which the smartwatch was paired. The back-end of the smartwatch app discards the old measurements unsuccessfully sent to the server as a preventive measure to avoid running out of memory because of an overflow of the internal buffers implemented to temporarily store the sensed measurements before being transmitted. As the 4G connection might slow down the data transmission, the amount of measurements buffered might be larger and, therefore, prone to losses. This was the reason why the measurements received from each activity occasionally contained discontinuities. As a pre-processing stage and to ensure the continuous stream of information, we trimmed the received data into segments of at least 20 s of consecutive sensor measurements. These segments are then used to populate the dataset.

This first version of the harAGE dataset contains 17 h 11 min 46 s of data from 19 participants (9 f, 10 m), with a mean age of 41.73 years and a standard deviation of 7.97 years. Before the data collection, participants read and signed an *Informed Consent Form* (ICF), which was previously approved by the competent ethics committee. A summary of the different activities considered in the dataset, and the amount of data available for each activity is provided in **Table 1**. Some participants partially completed the activities included in the protocol because of data transmission issues, or the impossibility to get access to a bike for the cycling-related activities.

4. METHODOLOGY

This section presents the methodology followed in this work to train end-to-end models for HAR using multimodal smartwatch data (cf. **Figure 2**). Section 4.1 describes the pre-processing applied to the raw measurements, Section 4.2 introduces the models implemented, and Section 4.3 summarizes their training details.

4.1. Data Preparation

In this passage, we describe the pre-processing applied to the raw measurements, which is different for each modality. After the pre-processing stage, the resulting information from each modality is segmented using windows of 20 s length and a 50 % overlap or without overlap, depending on whether the data is used for training or testing purposes, respectively. While each window contains 20 data points for the pedometer and the heart rate measurements, it contains 500 data points for the accelerometer measurements. The pre-processing applied to each modality is described below.

4.1.1. Pedometer Measurements

The Garmin Vivoactive 3 device allows retrieving the number of steps performed by the user since midnight. The absolute number of steps is not a suitable feature to model the current human activity, as the cumulative effect caused by the nature of the embedded sensor conditions the measurements. For instance, a high number of steps does not necessarily mean that the user is currently exercising, as the physical activity might have taken place a while—or even a long—ago. Instead, we hypothesize that the first and the second order derivatives computed from

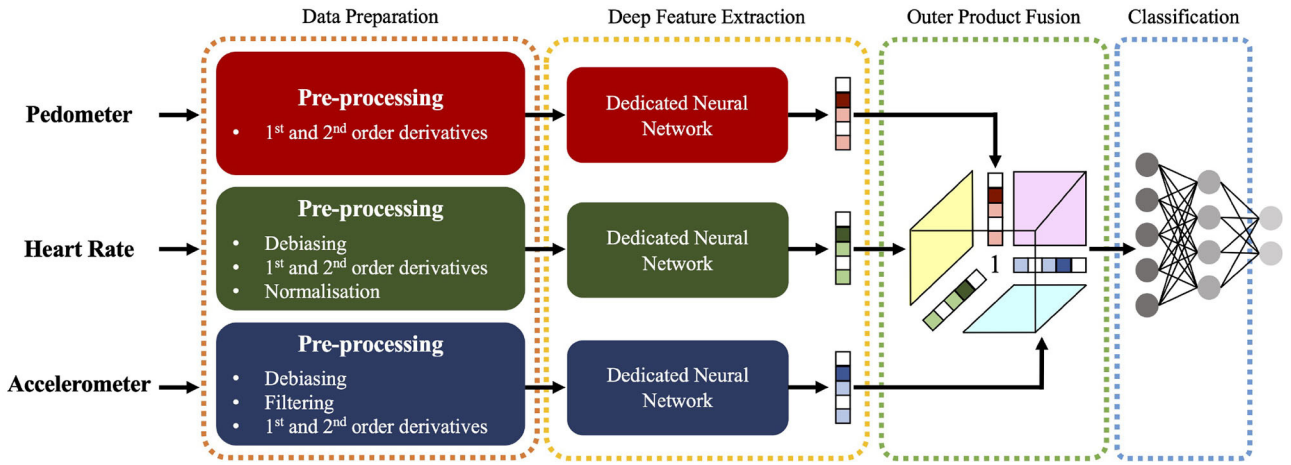

FIGURE 2 | Block diagram illustrating the system implemented, which receives the measurements from the pedometer, heart rate, and accelerometer modalities, and outputs the probability of the current sample to belong to the different activities considered.

the absolute number of steps could be more suitable features to characterize the users' activities, as these could model the velocity and the acceleration of the users' steps instantaneously. Hence, in our experiments, we use the first and the second order derivatives of the pedometer information as the features to extract from this modality.

4.1.2. Heart Rate Measurements

The characteristics of the human heart while exercising are person-dependent, as they might depend on a wide range of variables, including age, physical condition, or existing pathologies, among others. To remove this personal bias from our data, we compute the median of the heart rates collected from each participant during the resting activity individually and use this measurement as the personal, baseline heart rate. We opt for computing the median to avoid considering the outliers in the raw measurements. The heart rate measurements collected from all the activities performed by each participant are debiased using the corresponding personal, baseline heart rate. For this modality, we also compute the first and the second order derivatives of the debiased heart rate signals in order to better characterize their dynamics over time. Finally, we normalize the debiased heart rate signal by a factor of 220 BPM (beats per minute), which is widely considered as the maximum heart rate of a human being. Although the maximum heart rate is age-dependent from a theoretical point of view (Fox and Naughton, 1972), we disregard this factor and apply the same normalization parameter to all participants in the dataset. Therefore, in our experiments, we use the first and the second order derivatives of the debiased heart rate signals, and their normalized representation as the features to extract from this modality.

4.1.3. Accelerometer Measurements

It is sometimes possible to identify a person just by the way how she or he walks or moves. This observation leads us to hypothesize that the accelerometer measurements collected using a smartwatch can contain personal information that might interfere in the intrinsic movements of the activities considered in the harAGE dataset. To overcome this issue, we first read all the accelerometer measurements available in the dataset for each individual user separately and compute the median of the measurements in the x-, y-, and z-axes. We then use this information to debias the raw accelerometer measurements in a personalized manner. A 1-dimensional Gaussian filter, using a Gaussian kernel with a standard deviation of 1, is used to remove noises and smooth the accelerometer measurements in the 3 different axes separately (Zhuang and Xue, 2019). We finally compute the first and the second order derivatives of the debiased, filtered accelerometer measurements in order to better characterize the dynamics of this modality over time. Thus, in our experiments, we use the debiased, filtered accelerometer measurements and their first and second order derivatives as the features to extract from each axis of this modality.

4.2. Models Descriptions

The end-to-end models implemented in this work are composed of three different blocks: (i) the first block extracts dedicated deep learnt representations from the modality-dependent sequence of features defined in Section 4.1 (cf. Section 4.2.1), (ii) the second block, which is enabled when training multimodal models only, is in charge of fusing the embedded representations of the modalities selected (cf. Section 4.2.2), and (iii) the third and final block is responsible for performing the actual classification (cf. Section 4.2.3).

4.2.1. Deep Features Extraction

This block in the architecture is modality-specific; i. e., a dedicated feature extraction block processes the sequences of features from each modality separately. We compare two different network architectures for this task: an RNN, and a CNN. We use RNNs and CNNs as deep feature extractors,

TABLE 2 | Summary of the descriptive statistics (μ: mean, σ: standard deviation) computed from the UAR scores obtained when assessing the unimodal and the multimodal binary classification-based end-to-end models using nested LOSO-CV.

UAR [%]	RNN		CNN	
	μ	σ	μ	σ
f_{steps}	88.6	9.7	88.4	9.2
f_{hr}	82.8	15.2	85.2	12.4
f_{xyz}	70.2	19.3	90.9	13.5
$f_{steps\oplus hr}$	91.0	13.0	94.7	8.4
$f_{steps\otimes hr}$	92.9	7.2	91.8	13.7
$f_{steps\oplus xyz}$	89.3	13.2	91.5	13.2
$f_{steps\otimes xyz}$	88.6	12.8	88.3	11.4
$f_{hr\oplus xyz}$	85.8	14.4	95.2	8.3
$f_{hr\otimes xyz}$	86.6	14.1	92.6	13.2
$f_{steps\oplus hr\oplus xyz}$	**93.7**	6.4	**95.6**	5.5
$f_{steps\otimes hr\otimes xyz}$	90.5	15.1	95.4	4.6

The results compare the use of an RNN-based and a CNN-based architecture to extract deep learnt representations from the input modalities, and the fusion of the embedded representations in the multimodal models using a concatenation-based (represented with ⊕) and an outer product-based (represented with ⊗) approach. The bold values highlight the best results using each architecture.

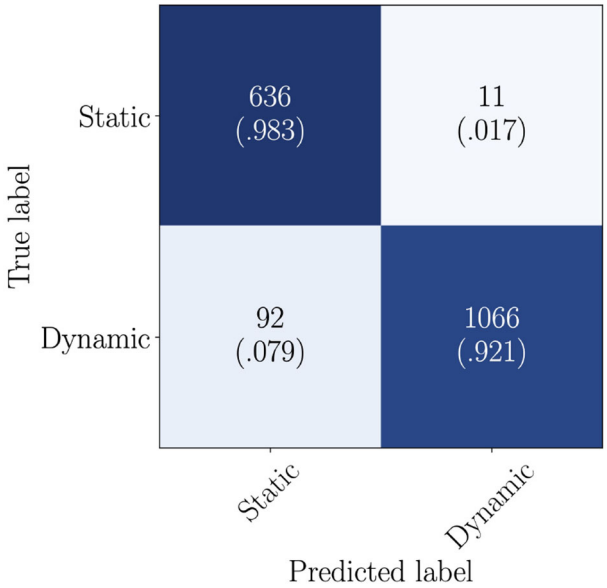

FIGURE 3 | Confusion matrix computed from the predictions inferred using the RNN-based approach exploiting $f_{steps\oplus hr\oplus xyz}$ as input features. The static class includes the lying, sitting, and standing activities, while the dynamic class includes the walking, running, stairs climbing, and cycling activities. Each cell contains the absolute and the relative number of samples of the actual class (row) classified into each possible activity of the recognition set (column). The relative information is also depicted using a color-scale: dark colors illustrate high percentages, while light colors, low percentages.

since they have been extensively used in the literature for such purpose. The RNN implements a single layer, bidirectional *Gated Recurrent Unit-Recurrent Neural Network* (GRU-RNN) with 8 hidden units. The CNN implements a single 1-dimensional convolutional layer with 8 filters, a kernel size of 2, and a stride of 1. Following this convolutional layer, we use 1-dimensional batch normalization, and the output is transformed using a *Rectified Linear Unit* (ReLU) function. A 1-dimensional adaptive average pooling layer is implemented at the end of this convolutional block, so it produces 2 features per filter. The parameters of the RNN- and the CNN-based architectures are designed, so they both produce 16 deep learnt features at the output. This way, we can fairly compare the performances between both approaches. The dimensionality of the deep learnt features is also engineered, so the resulting embeddings from the outer product-based fusion when training the multimodal models have a reasonable dimensionality in terms of computational cost.

4.2.2. Multimodal Fusion

One of the goals of this work is to investigate the suitability of using an outer product-based approach to fuse the embedded presentations learnt from different modalities in the problem of HAR. As a baseline, we use the simplest fusion method: the inner concatenation of the deep learnt representations from each modality. Representing these embedded representations for the pedometer, heart rate, and accelerometer modalities as f_{steps}, f_{hr}, and f_{xyz}, respectively, we mathematically define the concatenation-based fusion as:

TABLE 3 | Summary of the descriptive statistics (μ: mean, σ: standard deviation) computed from the UAR scores obtained when assessing the unimodal and the multimodal standard HAR-based end-to-end models using nested LOSO-CV.

UAR [%]	RNN		CNN	
	μ	σ	μ	σ
f_{steps}	30.2	11.1	30.7	5.6
f_{hr}	32.9	5.3	34.6	5.9
f_{xyz}	40.6	16.5	56.9	14.9
$f_{steps\oplus hr}$	51.4	10.1	47.7	13.7
$f_{steps\otimes hr}$	55.8	12.5	50.7	12.8
$f_{steps\oplus xyz}$	53.7	16.1	52.3	10.7
$f_{steps\otimes xyz}$	58.2	11.4	58.2	7.3
$f_{hr\oplus xyz}$	59.9	15.6	62.8	16.8
$f_{hr\otimes xyz}$	**62.5**	12.6	68.3	13.7
$f_{steps\oplus hr\oplus xyz}$	59.7	14.8	**69.5**	14.4
$f_{steps\otimes hr\otimes xyz}$	57.0	11.3	**69.5**	12.2

The results compare the use of an RNN-based and a CNN-based architecture to extract deep learnt representations from the input modalities, and the fusion of the embedded representations in the multimodal models using a concatenation-based (represented with ⊕) and an outer product-based (represented with ⊗) approach. The bold values highlight the best results using each architecture.

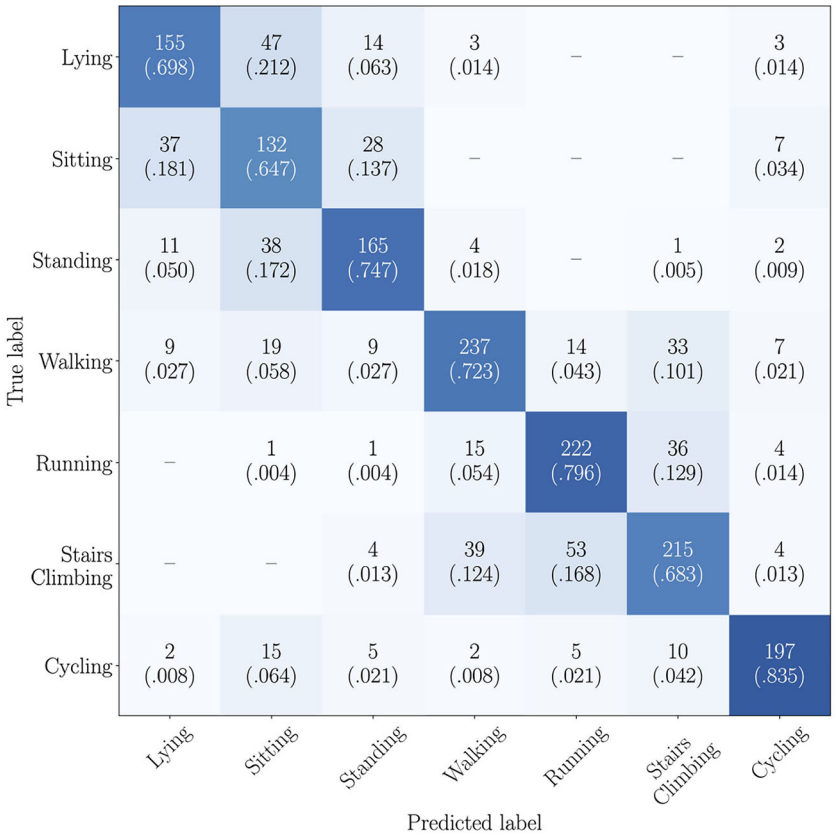

FIGURE 4 | Confusion matrix computed from the predictions inferred using the CNN-based approach exploiting $f_{steps \otimes hr \otimes xyz}$ as input features. Each cell contains the absolute and the relative number of samples of the actual class (row) classified into each possible activity of the recognition set (column). The relative information is also depicted using a color-scale: dark colors illustrate high percentages, while light colors, low percentages. Empty cells indicate no samples from the actual activity are classified into the corresponding class.

$$f_{steps \oplus hr \oplus xyz} = \begin{bmatrix} f_{steps} \\ f_{hr} \\ f_{xyz} \end{bmatrix}. \quad (1)$$

The dimensionality of the resulting embedded representation from the concatenation-based fusion is $\mathbb{R}^{16 \times m}$, where m indicates the number of modalities to be fused. When all three modalities are fused together ($m = 3$), the resulting embedded representation is $\in \mathbb{R}^{48}$. The outer product-based fusion proposed is inspired by the tensor fusion layer presented by Zadeh et al. (2017) and can be mathematically defined as:

$$f_{steps \otimes hr \otimes xyz} = \begin{bmatrix} f_{steps} \\ 1 \end{bmatrix} \otimes \begin{bmatrix} f_{hr} \\ 1 \end{bmatrix} \otimes \begin{bmatrix} f_{xyz} \\ 1 \end{bmatrix}. \quad (2)$$

When the three modalities are fused together, the outer product generates a cube with the following properties: (i) the original representations are preserved in the edges of the cube, (ii) each face of the cube contains information from the fusion of two modalities, and (iii) the inner part of the cube fuses information from the three modalities all together. The fused representation is flattened before being fed into the final, classification block of the network. The dimensionality of the resulting embedded representation from the outer product-based fusion is $\mathbb{R}^{(16+1)^m}$. When all three modalities are fused together ($m = 3$), the resulting embedded representation is $\in \mathbb{R}^{4913}$.

4.2.3. Classification

The classification block of the network implements two fully connected layers, preceded by a dropout layer with probability 0.3. The number of input neurons in the first fully connected layer depends on the number of modalities to be fused during the training process. The output of this layer produces a 16-dimensional representation, which is transformed using a ReLU activation function. This transformed representation is fed into the second fully connected layer, which contains as many neurons at the output as activities we need to classify our samples into, and uses a Softmax activation function. This way, the network outputs can be interpreted as probability scores.

4.3. Networks Training

For a fair comparison among the models, these are all trained under the exact same conditions. The pseudorandom number

TABLE 4 | Summary of the descriptive statistics (μ: mean, σ: standard deviation) computed from the UAR scores obtained when assessing the unimodal and the multimodal multi-class harAGE-based end-to-end models using nested LOSO-CV.

UAR [%]	RNN		CNN	
	μ	σ	μ	σ
f_{steps}	21.1	7.7	20.7	8.3
f_{hr}	23.6	3.9	24.1	4.7
f_{xyz}	34.3	11.7	43.4	12.5
$f_{steps \oplus hr}$	38.2	9.1	33.2	9.1
$f_{steps \otimes hr}$	38.5	9.2	35.8	9.1
$f_{steps \oplus xyz}$	46.5	10.4	41.3	11.1
$f_{steps \otimes xyz}$	47.4	10.5	51.0	12.5
$f_{hr \oplus xyz}$	46.3	12.9	46.8	11.6
$f_{hr \otimes xyz}$	48.7	12.9	60.5	13.6
$f_{steps \oplus hr \oplus xyz}$	**56.6**	10.1	54.8	13.3
$f_{steps \otimes hr \otimes xyz}$	49.7	9.6	**60.8**	12.8

The results compare the use of an RNN-based and a CNN-based architecture to extract deep learnt representations from the input modalities, and the fusion of the embedded representations in the multimodal models using a concatenation-based (represented with $\oplus$) and an outer product-based (represented with $\otimes$) approach. The bold values highlight the best results using each architecture.

generator is seeded at the initialization of the models for reproducibility purposes. The networks are trained to minimize the Categorical Cross-Entropy Loss, using Adam as the optimizer with a fixed learning rate of 10^{-3}. The metric selected to compare the inferred and the ground truth information is the *Unweighted Average Recall* (UAR). This metric allows us to account for the potential imbalance of the windowed sequences of data generated for the different activities. Hence, we define (1 − UAR) as the validation error to monitor the training progress. Network parameters are updated in batches of 64 samples and trained during a maximum of 150 epochs. We implement an early stopping mechanism to stop training when the validation error does not improve for 20 consecutive epochs. To assess the models, we follow a nested *Leave-One-Subject-Out Cross-Validation* (LOSO-CV) approach, splitting the data in the inner loop into 5 participant-independent folds. Each fold in the inner loop is trained during a specific number of epochs. Therefore, when modeling all the training material in the outer loop and to prevent overfitting, the training epochs are determined by computing the median of the training epochs processed in each fold. The resulting model is tested on the initially excluded participant. In compliance with the LOSO-CV approach, we apply this routine recursively, so each participant in the dataset can be used to test the performance of the trained models.

5. EXPERIMENTAL RESULTS

This section summarizes the experiments performed in this work and analyzes the results obtained. The resting activity is excluded from our experiments as, from a conceptual point of view, it can overlap with the lying, sitting, and standing activities. Nevertheless, the information collected during the resting activity is used in the context of our study to compute the personal, baseline heart rate (cf. Section 4.1.2). The pedometer information from 3 participants included in this first version of the harAGE dataset is corrupted. Consequently, we exclude all the data from these participants to train the models object of this study. We assess the performance of the models described in Section 4.2 from three different perspectives. Section 5.1 addresses the task as a binary classification problem. For this, we cluster the original activities into those that are static and those that are dynamic. We exclude the samples corresponding to the washing hands, strength workout, and flexibility workout activities. Section 5.2 tackles the recognition of the standard HAR dataset. In this case, we aim to model the lying, sitting, standing, walking, running, stairs climbing, and cycling activities and, therefore, we formulate the task as a 7-class classification problem. Finally, Section 5.3 addresses the task as a 10-class classification problem, targeting the automatic recognition of the whole set of activities considered in the harAGE dataset. Model performances are assessed by computing the UAR between the inferred and the ground truth annotations.

5.1. Binary Classification

The results obtained when tackling the task as a binary classification problem are summarized in **Table 2**. Analyzing the results, we observe that the multimodal models improve the performance of the unimodal models in most of the cases investigated. Comparing the performance of the multimodal models using the concatenation-based and the outer product-based approaches, the results indicate the suitability of the concatenation-based approach in this context, as it outperforms the outer product-based approach in 6 out of the 8 scenarios compared. When using the RNN-based architecture to extract deep learnt representations from the input modalities, the best UAR of 93.7 % is obtained with the model exploiting the pedometer, the heart rate, and the accelerometer modalities fused using the concatenation-based approach. The highest UAR of 95.6 % is achieved by the CNN-based architecture exploiting the three modalities together fused using the concatenation-based approach. The confusion matrix computed by comparing the activities inferred by this model and the ground truth annotations is depicted in **Figure 3**.

5.2. Standard HAR Classification

The results obtained when tackling the task as a 7-class classification problem are summarized in **Table 3**. The first observation of the results allows us to state that the multimodal models outperform the unimodal models in most of the cases investigated. The multimodal models using the RNN-based architecture to extract deep learnt representations from the input modalities and fusing the embedded information with the outer product-based approach surpass the models using the concatenation-based fusion in 3 out of the 4 scenarios compared. The multimodal models using the CNN-based architecture and

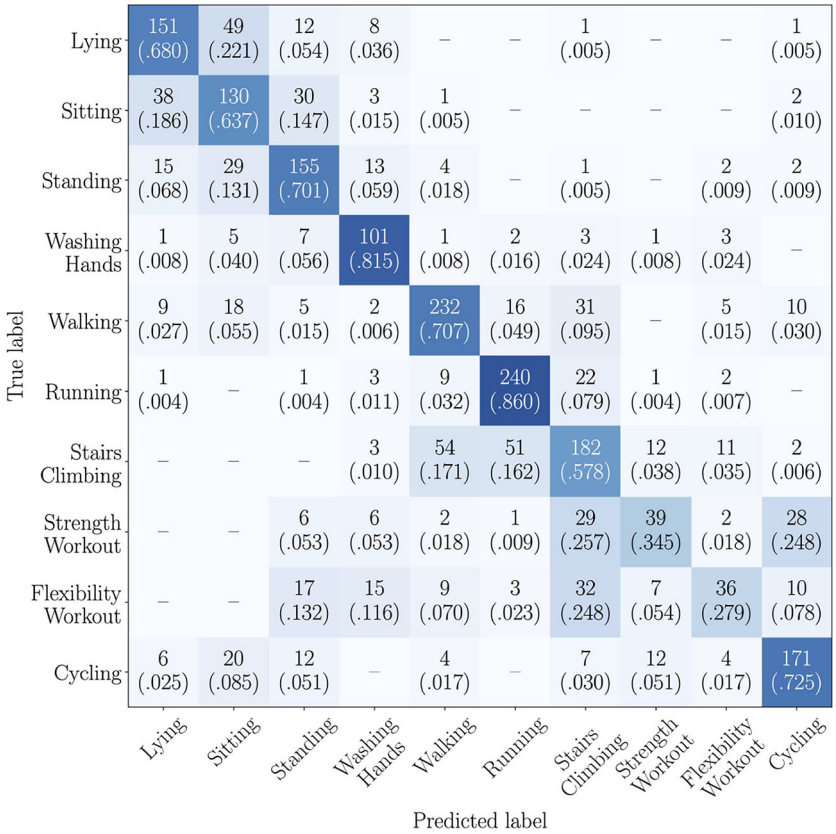

FIGURE 5 | Confusion matrix computed from the predictions inferred using the CNN-based approach exploiting $\boldsymbol{f}_{steps \otimes hr \otimes xyz}$ as input features. Each cell contains the absolute and the relative number of samples of the actual class (row) classified into each possible activity of the recognition set (column). The relative information is also depicted using a color-scale: dark colors illustrate high percentages, while light colors, low percentages. Empty cells indicate no samples from the actual activity are classified into the corresponding class.

fusing the embedded representations with the outer product-based approach improve the performance of the concatenation-based fusion in 4 out of the 4 cases compared. Although the $\boldsymbol{f}_{steps \oplus hr \oplus xyz}$ and the $\boldsymbol{f}_{steps \otimes hr \otimes xyz}$ models obtain the same mean from the individual UAR scores, the variance associated to the latter is lower. Hence, we consider the $\boldsymbol{f}_{steps \otimes hr \otimes xyz}$ model as the better of the two. The model using the RNN-based architecture with the highest UAR score of 62.5 % fuses the heart rate, and the accelerometer modalities with the outer product-based approach. The model with the highest UAR of 69.5 % implements the CNN-based architecture and fuses the pedometer, the heart rate, and the accelerometer modalities with the outer product-based approach. The confusion matrix computed by comparing the activities inferred by this model and the ground truth annotations is depicted in **Figure 4**.

5.3. Multi-Class harAGE Classification

The results obtained when tackling the task as a 10-class classification problem are summarized in **Table 4**. The results obtained indicate that the multimodal models surpass the unimodal models in most of the cases. From the results, we also observe that the multimodal models fusing the embedded representations with the outer product-based approach outperform the concatenation-based fusion in all the cases investigated with one exception: the multimodal RNN-based network fusing the pedometer, heart rate, and accelerometer information using the concatenation-based approach surpasses the outer product-based fusion. The model with the best performance using the RNN-based architecture scores a UAR of 56.6 %, exploiting the pedometer, the heart rate, and the accelerometer modalities fused with the concatenation-based approach. The highest UAR of 60.8 % is obtained with the CNN-based model that fuses the pedometer, the heart rate, and the accelerometer modalities using the outer product-based approach. The confusion matrix computed by comparing the activities inferred by this model and the ground truth annotations is depicted in **Figure 5**. As it can be seen in the confusion matrix, the strength and the flexibility workout activities are the most difficult ones to be recognized and cause the highest confusion. While the samples corresponding to the strength workout activities tend to be misclassified into the stairs climbing and the cycling activities, the samples corresponding to the flexibility workout activities tend to be mainly misclassified into the stairs climbing activity.

6. CONCLUSIONS

This work focused on the use of an outer product-based approach to fuse the embedded representations learnt from the pedometer, the heart rate, and the accelerometer information collected using a smartwatch for the problem of HAR. The best results obtained when tackling the task as a 2-class, 7-class, and 10-class classification problem were achieved with the multimodal models using a CNN-based architecture to extract deep learnt representations from the pedometer, the heart rate, and the accelerometer modalities as input data. The outer product-based fusion obtained the highest UAR scores in the 7-class, and the 10-class problems, and ranked the second highest UAR score in the 2-class problem. These results supported the suitability of fusing the pedometer, the heart rate, and the accelerometer information with the proposed outer product-based approach for the task at hand.

The pre-processing applied to the accelerometer measurements is one of the limitations of this work. As described in Section 4.1.3, we computed the personal, debiasing parameters for the accelerometer measurements using all the data available from the current participant in the dataset. In a real-life scenario, these parameters should be computed and updated on the fly. In terms of model performance, we expect the trained models to underperform when a new user uses the system for the first time, and improve as the user keeps using the system, once the personal, debiasing accelerometer parameters stabilize.

Further research directions include the investigation of other techniques to fuse the information from the available modalities. Additionally, exploring whether the high performance of the binary classification problem can be used to improve the performance of the N-class classification problems—with $N > 2$—in a multi-task or a transfer learning set up might also be worth researching.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Institute of Computer Science, Foundation of Research and Technology—Hellas, Greece. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

AM-R, AS, MP, and BS conceptualized the study. AM-R ran the machine learning experiments. AS and MP engaged participants and collected data. AM-R and AS did literature analysis, manuscript preparation, and editing. All authors revised, read, and approved the final manuscript.

FUNDING

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 826506 (sustAGE).

ACKNOWLEDGMENTS

The authors would like to sincerely thank the participants who took part in the collection of the investigated dataset. We would also like to thank Georgios Athanassiou and Michalis Maniadakis for their contributions in the protocol design for collecting the data.

REFERENCES

Ahmed, N., Rafiq, J. I., and Islam, M. R. (2020). Enhanced human activity recognition based on smartphone sensor data using hybrid feature selection model. *Sensors* 20, 317. doi: 10.3390/s20010317

Altun, K., and Barshan, B. (2010). "Human activity recognition using inertial/magnetic sensor units," in *Proceedings of the International Workshop on Human Behavior Understanding* (Istanbul: Springer), 38–51.

Ashry, S., Ogawa, T., and Gomaa, W. (2020). CHARM-deep: continuous human activity recognition model based on deep neural network using IMU sensors of smartwatch. *Sensors* 20, 8757–8770. doi: 10.1109/JSEN.2020.2985374

Bayat, A., Pomplun, M., and Tran, D. A. (2014). A study on human activity recognition using accelerometer data from smartphones. *Procedia Comput. Sci.* 34, 450–457. doi: 10.1016/j.procs.2014.07.009

Chen, Y., and Shen, C. (2017). Performance analysis of smartphone-sensor behavior for human activity recognition. *IEEE Access* 5, 3095–3110. doi: 10.1109/ACCESS.2017.2676168

Fox, K. R. (1999). The influence of physical activity on mental well-being. *Public Health Nutr.* 2, 411–418. doi: 10.1017/s1368980099000567

Fox, S. M., and Naughton, J. P. (1972). Physical activity and the prevention of coronary heart disease. *Prevent. Med.* 1, 92–120.

Hassan, M. M., Uddin, M. Z., Mohamed, A., and Almogren, A. (2018). A robust human activity recognition system using smartphone sensors and deep learning. *Future Gen. Comput. Syst.* 81, 307–313. doi: 10.1016/j.future.2017.11.029

Holzinger, A., Malle, B., Saranti, A., and Pfeifer, B. (2021). Towards multi-modal causability with Graph Neural Networks enabling information fusion for explainable AI. *Inf. Fusion* 71, 28–37. doi: 10.1016/j.inffus.2021.01.008

Khaire, P., Kumar, P., and Imran, J. (2018). Combining CNN streams of RGB-D and skeletal data for human activity recognition. *Pattern Recognit. Lett.* 115, 107–116. doi: 10.1016/j.patrec.2018.04.035

Khan, A. M., Lee, Y.-K., Lee, S. Y., and Kim, T.-S. (2010). "Human activity recognition via an accelerometer-enabled-smartphone using kernel discriminant analysis," in *Proceedings of the 5th International Conference on Future Information Technology* (Busan: IEEE), 6.

Kim, J., Koh, J., Kim, Y., Choi, J., Hwang, Y., and Choi, J. W. (2018). "Robust deep multi-modal learning based on gated information fusion network," in *Proceedings of the Asian Conference on Computer Vision* (Perth, WA: Springer), 90–106.

Kwon, Y., Kang, K., and Bae, C. (2014). Unsupervised learning for human activity recognition using smartphone sensors. *Exp. Syst. Appl.* 41, 6067–6074. doi: 10.1016/j.eswa.2014.04.037

Lara, O. D., Pérez, A. J., Labrador, M. A., and Posada, J. D. (2012). Centinela: a human activity recognition system based on acceleration and vital sign data. *Pervasive Mobile Comput.* 8, 717–729. doi: 10.1016/j.pmcj.2011.06.004

Li, Q., Gravina, R., Li, Y., Alsamhi, S. H., Sun, F., and Fortino, G. (2020). Multi-user activity recognition: challenges and opportunities. *Inf. Fusion* 63, 121–135. doi: 10.1016/j.inffus.2020.06.004

Lin, W.-Y., Verma, V. K., Lee, M.-Y., and Lai, C.-S. (2018). Activity monitoring with a wrist-worn, accelerometer-based device. *Micromachines* 9, 450. doi: 10.3390/mi9090450

Mallol-Ragolta, A., Semertzidou, A., Pateraki, M., and Schuller, B. (2021). "harAGE: a novel multimodal smartwatch-based dataset for human activity recognition," in *Proceedings of the 16th International Conference on Automatic Face and Gesture Recognition* (Jodhpur: IEEE), 7.

Mekruksavanich, S., and Jitpattanakul, A. (2020). "Smartwatch-based human activity recognition using hybrid LSTM network," in *Proceedings of Sensors* (Rotterdam: IEEE), 4.

Penedo, F. J., and Dahn, J. R. (2005). Exercise and well-being: a review of mental and physical health benefits associated with physical activity. *Curr. Opin. Psychiatry* 18, 189–193. doi: 10.1097/00001504-200503000-00013

Prakash, A., Chitta, K., and Geiger, A. (2021). "Multi-modal fusion transformer for end-to-end autonomous driving," in *Proceedings of the Conference on Computer Vision and Pattern Recognition* (Nashville, TN: IEEE), 7077–7087.

Priyasad, D., Fernando, T., Denman, S., Sridharan, S., and Fookes, C. (2021). Memory based fusion for multi-modal deep learning. *Inf. Fusion* 67, 136–146. doi: 10.1016/j.inffus.2020.10.005

Qi, J., Wang, Z., Lin, X., and Li, C. (2018). Learning complex spatio-temporal configurations of body joints for online activity recognition. *IEEE Trans. Hum. Mach. Syst.* 48, 637–647. doi: 10.1109/THMS.2018.2850301

Ronao, C. A., and Cho, S.-B. (2014). "Human activity recognition using smartphone sensors with two-stage continuous hidden markov models," in *Proceedings of the 10th International Conference on Natural Computation* (Xiamen: IEEE), 681–686.

Ronao, C. A., and Cho, S.-B. (2016). Human activity recognition with smartphone sensors using deep learning neural networks. *Exp. Syst. Appl.* 59, 235–244. doi: 10.1016/j.eswa.2016.04.032

Shahmohammadi, F., Hosseini, A., King, C. E., and Sarrafzadeh, M. (2017). "Smartwatch based activity recognition using active learning," in *Proceedings of the International Conference on Connected Health: Applications, Systems and Engineering Technologies* (Philadelphia, PA: IEEE), 321–329.

Tapia, E. M., Intille, S. S., Haskell, W., Larson, K., Wright, J., King, A., et al. (2007). "Real-time recognition of physical activities and their intensities using wireless accelerometers and a heart rate monitor," in *Proceedings of the 11th International Symposium on Wearable Computers* (Boston, MA: IEEE), 4.

Wan, S., Qi, L., Xu, X., Tong, C., and Gu, Z. (2020). Deep learning models for real-time human activity recognition with smartphones. *Mobile Netw. Appl.* 25, 743–755. doi: 10.1007/s11036-019-01445-x

Weiss, G. M., Timko, J. L., Gallagher, C. M., Yoneda, K., and Schreiber, A. J. (2016). "Smartwatch-based activity recognition: a machine learning approach," in *Proceedings of the 3rd International Conference on Biomedical and Health Informatics* (Las Vegas, NV: IEEE), 426–429.

Wu, A., and Han, Y. (2018). "Multi-modal circulant fusion for video-to-language and backward," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence* (Stockholm: IJCAI), 1029–1035.

Zadeh, A., Chen, M., Poria, S., Cambria, E., and Morency, L.-P. (2017). "Tensor fusion network for multimodal sentiment analysis," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (Copenhagen: ACL), 1103–1114.

Zhuang, Z., and Xue, Y. (2019). Sport-related human activity detection and recognition using a smartwatch. *Sensors* 19, 21. doi: 10.3390/s19225001

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.